DOCUMENT RESUME

ED 390 942 TM 024 585

AUTHOR Linacre, John Michael
TITLE ANOVA with Rasch Measures.
PUB DATE Apr 95
NOTE 9p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, April 18-22, 1995).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Analysis of Variance; *Demography; *Error of Measurement; *Estimation (Mathematics); *Item Response Theory; *Scaling
IDENTIFIERS *Rasch Model

ABSTRACT

Various methods of estimating main effects from ordinal data are presented and contrasted. Problems discussed include: (1) at what level to accumulate ordinal data into linear measures; (2) how to maintain scaling across analyses; and (3) the inevitable confounding of within cell variance with measurement error. An example shows three methods of estimating demographic main effects from student responses to an arithmetic test previously reported and three approaches to reporting those results: (1) a routine Rasch analysis of the 776 students by 4 items; (2) accumulating the data at a main effect level and Rasch analysis; and (3) accumulating the data at a main cell level and Rasch analyzing. The usefulness of each of the methods is discussed. (Contains one figure and five references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
JOHN M. LINACRE
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)"

ED 390 942

ANOVA with Rasch Measures

By

John Michael Linacre
University of Chicago

Paper presented at the
American Educational Research Association Annual Meeting

San Francisco, California
April 1995

MESA Psychometric Laboratory
Department of Education
University of Chicago
5835 S. Kimbark Ave
Chicago IL 60637-1609

Tel: (312) 702-1596, FAX (312) 702-0248
E-mail: MESA@uchicago.edu

TM024585

BEST COPY AVAILABLE

ANOVA with Rasch Measures

Abstract:
Various methods of estimating main effects from ordinal data are presented and contrasted. Problems discussed include: 1) at what level to accumulate ordinal data into linear measures, 2) how to maintain scaling across analyses, 3) the inevitable confounding of within cell variance with measurement error. An example shows three methods of estimating demographic main effects from student responses to an arithmetic test, and three approaches to reporting those results.

Text:
Analysis of variance (ANOVA) encompasses a family of techniques for discovering what differences between sizes of observations could be systematic and what could be random. A simple example is an investigation of whether men are taller than women. A *t*-test could be formed by computing the difference between the mean height of a sample of men and the mean height of a sample of women, then dividing this difference by the standard deviation of all the heights from the grand mean. Yet problems arise even in this simple example: What if the sample sizes are different? What if the heights of the men are much more dispersed than the heights of the women? What if the distributions of heights are skewed? What if the heights are measured with different precisions? What if the "heights" are measured on a non-linear scale?

Empirical data is always so complex that any attempt at analysis of variance requires that simplifying assumptions be made. Frequent ones are that the numerical quantities to be analyzed are measures 1) located on a linear scale and 2) observed with perfect precision. With ordinal observations both these assumptions are untenable. For instance, for rating scales, each ordinal observation represents, not a point, but a zone of performance from a conceptually infinite range. Further, each nominal value represents a performance zone of a different size. The extreme observations represent infinite ranges above or below the other category zones. The sizes of the zones corresponding to the intermediate categories depend on how those zones are defined and used. Further, the choice of which category to report may be influenced in many minor irrelevant ways. Thus rating scale data are not linear and precise, but ordinal and diffuse.

Rasch analysis permits the analyst to address some ANOVA assumptions directly, but Rasch analysis also prompts the analyst to consider other assumptions of which the analyst may have been unaware.

Analysis with Rasch Measures

When ordinal data usefully fit the Rasch model, measures are estimated on a scale **constructed** to be linear. This contrasts with numbers that are on scales only **asserted** to be linear for the convenience of later analysis. Such assertion occurs when essay ratings and other clearly non-linear observations are subjected to linear-based analytic approaches, such as generalizability theory.

Example: A numerical value, the raw score, asserted to be the ability a person *n* measured on a linear scale with pin-point precision:

$$\sum_i X_{ni} \qquad (1)$$

where X_{ni} is the scored response of person n to item i.

Rasch analysis estimates **linear** measures and their standard errors. ANOVA based on linear estimates with standard errors is more demanding than analysis based on point estimates, but it can be readily performed with sophisticated statistical software, such as HLM.

Example: A numerical value, a logit measure, constructed to be the ability person *n* measured on a linear scale with estimable precision:

$$\hat{B}n \pm SE(\hat{B}n) \qquad (2)$$

where $\hat{B}_n$ is the measure of person n estimated from all items.

Obtaining measures from ordinal observations

What degree of data summary most usefully underlies Rasch measures? Estimation of measures requires replication in the data, but which replications should be summarized into measures? For instance, in order to estimate a subject's math ability, we could administer a 100 item test containing 10 subtests. Each subtest addresses one strand of math competence: addition, subtraction, multiplication, word problems, etc. Two analytical approaches spring to mind:

a) Estimate the subject measure directly from the 100 items, simultaneously with the measures of the N-1 other subjects. This approach can be performed directly with a Rasch analysis of an N person by 100 item response matrix. This procedure can be summarized as:

$$\sum_{i=1}^{100} X_{ni} \rightarrow \hat{B}_n \pm SE(\hat{B}_n) \qquad (3)$$

b) Estimate a subject sub-measure for each of the 10 subtests, and then combine these sub-measures to produce subject measures:

$$\{\sum_{i=1 \in S_t}^{10} X_{ni}\} \rightarrow \{\hat{B}_{nt} \pm SE(\hat{B}_{nt})\} \rightarrow \hat{B}_n \pm SE(\hat{B}_n) \qquad (4)$$

where S_t indicates subtest t, and B_{nt} indicates the ability measure of person n on subtest t.

These two methods, a) and b), usually produce similar, but not identical results. But even for similar results to be obtained, further constraints must be introduced. For instance, an immediate dilemma is presented by subjects who have extreme scores (zero or perfect) on one or more subtests, but not on the whole test. Such subjects will have poorly defined measures on some subtests. Either these subjects must be put to one side as inestimable or arbitrary (Bayesian) measures must be imputed for extreme subtest scores. In either case, the relationship between subtest and whole test measures is no longer indisputable.

Another complication in method b) is that independent analysis of each subtest permits each subtest to exhibit its own discrimination depending on the local stochasticity of the test. The subtests' logits have different "length" (Linacre & Wright 1989). Rescaling the dispersion of abilities to be uniform across subtests assists with this:

$$B_{nt} \rightarrow \frac{B_{nt}}{SD(B_{\{n\}t})} * \sqrt{\frac{\sum_t SD^2(B_{\{n\}t})}{\sum_t 1}} \qquad (5)$$

Further, in method b), routine analysis of each subtest allows each subtest to define its own local origin at the mean of the subtest's item difficulties. A solution would be to use common person equating of subtests to locate the person abilities within each subtest in one frame of reference that maintains the grand mean of all subtest ability measures:

$$B_{nt} \rightarrow B_{nt} - \bar{B}_{\{n\}t} + \bar{B}_{\{nt\}} \qquad (6)$$

Another difficulty in method b) is the differential precision of the subtest measures. A solution is to treat each subtest measure as though it were the result of a separate study and then apply the method of "effect sizes" used in meta-analysis:

$$(7)\ B_n = \frac{\sum_t \frac{B_{nt} - \bar{B}_{n\{t\}}}{SE(B_{nt})}}{\sum_t \frac{1}{SE(B_{nt})}} \pm SE(B_n)$$

For details of computing the combined standard error, see Hedges & Olkin (1985).

Measuring within a common frame of reference

The problems of differential subtest discrimination and origin can be addressed by using the "item bank" approach of establishing all subtest measures in one frame of reference. This method is successful when a subject's performance level across subtests is fairly homogeneous (Wright 1994).

1) Perform a joint analysis of all subtests in order to obtain item calibrations for all subtests in one common frame of reference. (If the item calibrations are already known, e.g., when tests are constructed from item banks, then this step is not needed):

$$\sum_{n} X_{ni} \rightarrow \hat{D}_i \qquad (8)$$

2) Analyze each subtest separately, but with the items anchored at their common calibrations:

$$\sum_{i \in t} X_{ni} \mid \{D_i\} \rightarrow \hat{B}_{nt} \qquad (9)$$

3) Obtain joint calibrations by applying the effect-size method, equation (7) above.

Accumulating at higher levels

In many instances, the focus of attention is not at the subject level, but at some higher level, e.g., classroom, school, school district, region or country. In these cases, it may be preferable to consider each student as, say, a random representative of a fixed school effect. Thus, in a particular school there may be 150 correct responses and 50 incorrect responses to item 10. It is the school that becomes the object of measurement with a score of 150/200. Many market surveys and political polls start off at this level, because individual response strings are not recorded, but responses are immediately accumulated by category. These accumulated responses are difficult to analyze with Rasch software designed to handle rectangular data matrices with one response per cell, but are straightforward with more general purpose Rasch software, such as Facets (Linacre 1987).

Conceptualizing and estimating higher level effects requires careful thought whatever type of data is to be aggregated, e.g., linear measures, correlations, or ordinal responses. A simple, but not trivial, example is based on data presented in the Facets manual and attributed to Mislevy.

In Mislevy's data, 776 students (black and white, males and females) respond to a four item arithmetic test. The researcher is interested in gender and race effects and interactions. Several analytical methods appear reasonable:

Method a) Perform a routine Rasch analysis of 776 students by 4 items.
This yields a measure and standard error for each student. Then identify each student by race and gender. Compute the main effects and their standard errors for each race and gender by combining effect sizes or by information weighting (see Linacre 1992). Figure 1 shows the student measures produced by method (a) and the results of a simple unweighted decomposition into main effects. Despite the potential voluminosity of this method, it is computationally simple with standard software. Nevertheless, there are snags:

a.1) Extreme scores. In Mislevy's data, 99 students succeeded on none of the items. 134 succeeded on all of them. The measures and standard errors imputed to these students will have an arbitrary aspect, but they will also be influential in the outcome of the analysis. Consequently, a different analysis that avoids this particular arbitrariness is preferable. Accumulating the ordinal data at the main effect level would prevent any occurrence of an extreme score.

a.2) Singular data. If the math test had consisted of only one item, then every score would have been extreme. Thus individual subject measures would be arbitrary, though accumulated main effect measures based on marginal scores would be meaningful. Each main effect's measure could be modelled as the outcome of a series of as many Bernoulli trials as there were subjects relevant to that main effect.

a.3) Latent traits only meaningful at higher levels. In the Mislevy data, the items are homogeneous enough that it is reasonable to think of a student as having the same ability level across all of them. But a similar study might consider 4 items with each item selected from a very different content area. In this case, it may not be reasonable

to think of a student having a particular constant level across items, but it may be reasonable to think of such a level for a school. Thus, if the items were to relate to participation in extra-curricular activities, each activity might be so different that it is unreasonable to think of each student as having one overall participation proclivity. But at the school level, some school environments will promote greater participation and some less. This school-level effect is reflected in school level data analysis, rather than individual analysis.

Method b) Accumulate the data at a main effect level and Rasch analyze.
For the Mislevy data, a promising model is

$$\log\left(\frac{P_{rgi}}{P_{rg0}}\right) = R_r + G_g - D_i \qquad (10)$$

where R_r is the main effect for Race r, G_g is the main effect for Gender g. Results based on this model are shown in Figure 1.

With this model, there is no need for a statistically balanced design. Each race and gender can appear as often as is convenient for substantive, e.g., demographic, reasons. If, say, there are more males than females in the data set, then the male main effect will be estimated more precisely than the female effect. Further, the occurrence of extreme scores is very unlikely in any reasonably sized data set. But again there are predicaments:

b.1) Fixed effect vs. Random effect. Equation (10) models each main effect as a fixed effect, as though each member of each gender would exhibit the same measure except for measurement error. In fact, we expect members of each gender to differ in measure. We would find it more convenient to think of our subjects as normally distributed, with a particular mean and variance, i.e., as exhibiting a random effect. But the random effect variance interacts with the probabilistic nature of the Rasch model, and so cannot be conveniently parameterized. Nevertheless, differential main effect variances are reflected in mean-square fit statistics. The more variance among the subjects producing a main effect relative to the other main effect variances, the larger the mean-square.

b.2) Error variance vs. within variance. Whenever separate observations are to be represented by one parameter, the Rasch model specifies that all between observation variance, not otherwise parameterized, is to be explained by the probabilistic aspect of the measurement model. Thus if 50 book-keepers with high arithmetic skills are accumulated with 50 first graders with low arithmetic skills and all are to be represented by one parameter, then their joint "fixed effect" will correspond to a raw score of 50/100. The very real within-group variance is combined with the inevitable imprecision in the observations and together they are modelled as measurement error. Of course, this combining of within variance and error variance occurs even at the individual subject level, because no one performs at exactly the same level across all items of a test. In the individual subject case, however, it is usually reasonable (and often necessary) to think of the subject's level as steady.

b.3) Change of measurement scale. A side-effect of combining error variance and within variance is redefinition of the measurement scale unit. The logit is defined in terms of the variance in the observations. As the apparently random variance (error and within) connected with each parameter increases, the less discriminating the measurement system becomes. Then estimates of the same pair of parameters become closer together in logit terms. The difference between the mean abilities of each gender, computed according to more discriminating method a) above, will be larger than according to less discriminating method b). The effect of change of measurement scale can be diminished in two ways:

b.3.1) Maintain the frame of reference by pre-calibrating the items - see "Measuring with a common frame of reference" (above). The calibrations can either be obtained by method a) and then anchored in method b) or *vice-versa*. Often, particularly with rating scales or very sparse data, one method produces much more stable, useful and defensible item and rating scale calibrations. Nevertheless, the greater the within variance, the less comparable the results become. Mean-square statistics reported, on average, far away from 1.0 and large logit displacements are an indication that pre-calibrated values are not working well in the target analysis. In Figure 1, anchoring is seen to stretch the scale only a small amount for Mislevy's data set.

b.3.2) Rescale to a common linear scale. Though the logit has a well-defined probabilistic interpretation, this is often irrelevant to the final purposes of the analysis. Consequently, it may be more meaningful to equate the results

of method a) and method b) by asserting common means and variances for comparable parts of the analyses. Thus the results of routine, unanchored analyses according to method a) and b) could be made comparable by rescaling both outputs so that the mean and standard deviations of the item difficulties become identical. This would maintain the substantive meaning of the linear scale, but lose much of the probabilistic interpretation. In Figure 1, rescaling item calibrations from method b) to match those from method a) increases the distances between main effects beyond those in method a).

b.4) Loss of diagnostic power. After subject responses have been accumulated, it is no longer possible to identify subjects with uneven response strings. Further, the entire response string of very high or very low performing subjects may be flagged as uncharacteristic of the group, and so misfitting. Consequently, it is often advisable to perform data validation at as low a data level accumulation level as convenient, e.g., at subject or subtest level.

b.5) Interactions. Once the main effects have been estimated, interaction effects can be computed:

$$\log\left(\frac{P_{rgi}}{P_{rg0}}\right) = \hat{R}_r + \hat{G}_g - D_i + RG_{rg} \qquad (11)$$

where RG_{rg} is the effect due to the interaction of Race r with Gender g. These interaction effects are measures in the same frame of reference as the main effects. They can, however, be difficult to interpret, because the reported interaction measure adds to the sum of the corresponding main effects.

Method c) Accumulate the data at a main cell level and Rasch analyze.
Since, for the Mislevy data, interaction effects are of the same order of magnitude as main effects, within parameter variance is probably of the same size as between parameter variance. This suggests that a more effective measurement model might be:

$$\log\left(\frac{P_{rgi}}{P_{rg0}}\right) = RG_{rg} - D_i \qquad (12)$$

Results based on this model are shown in Figure 1. For these data, this model has one more unconstrained demographic parameter than the model shown for method b) in equation (10), but the model fit is about the same. The reparameterization has redistributed the apparent error variance by making some within-cell variance into between-cell variance and *vice versa*. Consequently, the measurement system has about the same discrimination, i.e., the logit distances between equivalent parameter estimates (item difficulties) is about the same. On the other hand, the interaction effects of equation (11) are incorporated into the main effects, making the output straightforward to interpret. Anchored and rescaled results using this model are also shown in Figure 1, and are similar to those obtained with the model in method b).

Conclusion

The problems in data analysis, as presented in this paper, are not the result of using the Rasch model. They have existed all along, though usually hidden behind the apparently unequivocal, linear form of the data and the deceptively obvious way to analyze it. In fact, ordinal data is always blurry and non-linear. Moreover, there are many ways in which it can be analyzed. It is the task of the analyst to discover meaningful ways to analyze the data and to communicate the results of those analyzes to the target audience. Some have been suggested in this paper.

Hedges LV & Olkin I. 1985. Statistical methods for meta-analysis. New York: Academic Press.

Linacre JM. 1987. Facets computer program for man-facet Rasch measurement. Chicago: MESA Press.

Linacre JM. 1992. Treatment effects. Rasch Measurement Transaction 8:2 218-219.

Linacre JM & Wright BD. 1989. The "length" of a logit. Rasch Measurement Transactions 3:2 p. 53-55.

Wright BD. 1994. Part-test vs. whole-test measures. Rasch Measurement Transactions 8:3 p. 376-377.

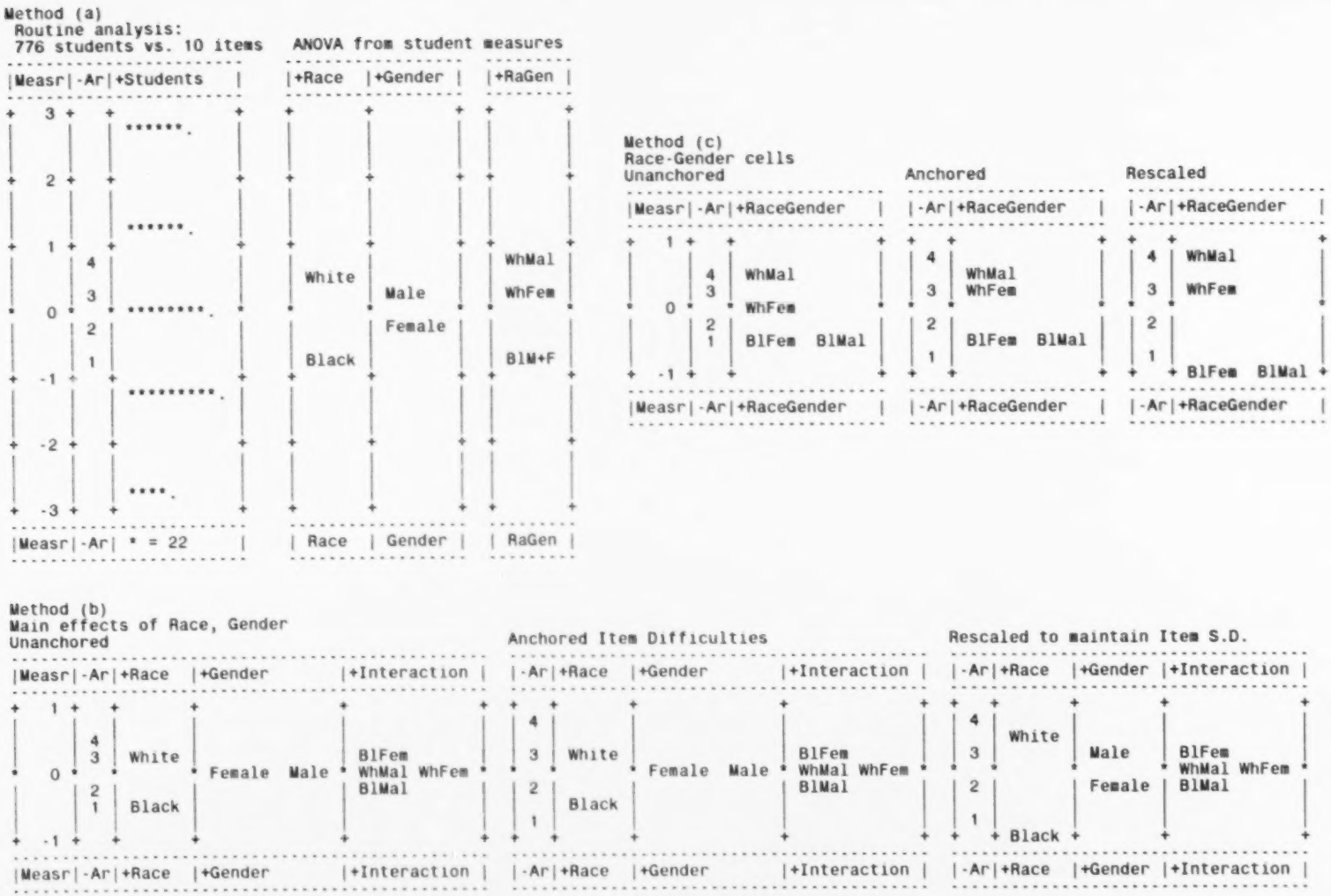

Figure 1. Results of 3 methods for analyzing Mislevy data.

BEST COPY AVAILABLE